The Joint European Compound Library

Besnard, Jérémy; Jones, Philip S.; Hopkins, Andrew L.; Pannifer, Andrew D.

Published in:
Drug Discovery Today

DOI:
10.1016/j.drudis.2014.08.014

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

ELSEVIER

feature

The Joint European Compound Library: boosting precompetitive research

Jérémy Besnard[1], Philip S. Jones[2], Andrew L. Hopkins[1] and Andrew D. Pannifer[2],
a.pannifer@dundee.ac.uk

The Joint European Compound Library (JECL) is a new high-throughput screening collection aimed at driving precompetitive drug discovery and target validation. The JECL has been established with a core of over 321 000 compounds from the proprietary collections of seven pharmaceutical companies and will expand to around 500 000 compounds. Here, we analyse the physicochemical profile and chemical diversity of the core collection, showing that the collection is diverse and has a broad spectrum of predicted biological activity. We also describe a model for sharing compound information from multiple proprietary collections, enabling diversity and quality analysis without disclosing structures. The JECL is available for screening at no cost to European academic laboratories and SMEs through the IMI European Lead Factory (http://www.europeanleadfactory.eu/).

Background

The decline in the rate of new molecular entity (NME) discovery is well documented [1] and the cost of discovering a drug has risen dramatically over the past 20 years. Boosting precompetitive research has become a major theme to reverse this trend by sharing risk in early-stage research [2,3] and by bringing together key areas of expertise from diverse organisations. These precompetitive initiatives have generated highly successful partnerships such as the Structural Genomics Consortium and the Human Blood Plasma Metabolome Consortium, created to build technology platforms supporting drug design and development of biomarkers. However, access by external organisations to the pharmaceutical companies' proprietary screening collections has, until now, been very limited owing to their high intellectual property value. The compounds embody a large amount of in-house molecular design expertise and synthetic effort, and represent a primary source of future drugs. These very qualities make the libraries an extremely valuable resource for academic groups seeking to identify inhibitors and tool compounds for their biological targets. Making compounds from these proprietary libraries available to academic laboratories and small companies has therefore the potential to expand the pool of pharmacologically validated drug targets available to the industry and to identify new lead compounds.

The Joint European Compound Library (JECL) is a key component of the IMI European Lead Factory (ELF) and brings together over 321 000 high-quality, drug-like and lead-like compounds from the in-house collections of seven large pharmaceutical companies into a single screening collection. Over the course of the next four years, the JECL will be expanded through bespoke synthesis of 200 000 compounds to address regions of chemical space unexplored or underrepresented by the core set. These additional compounds are being designed by industry and academic scientists and synthesised by companies across Europe. This collection is now available to the academic community and to biotech companies across Europe to be screened at the ELF (http://www.europeanleadfactory.eu/). The ELF provides established expertise in compound logistics, high-throughput screening and characterisation infrastructure, and medicinal chemistry for triaging the screen output. The JECL is also being screened by the contributing pharmaceutical companies to extend their existing hit identification campaigns. In this article, we describe the physicochemical

http://dx.doi.org/10.1016/j.drudis.2014.08.014

profile of the collection, chemical diversity and pharmacological profile of the core 321 000 compounds currently available for screening. We also describe a model for assessing diversity of compounds from multiple collections while allowing compound structures to remain confidential.

Design of the core collection

A key consideration in combining compound libraries from different pharmaceutical companies is the degree of overlap between them. As a result of mergers in the pharmaceutical industry and consequent merging of screening collections, it has been possible to gain an insight into the similarities of proprietary screening collections. In two published cases [4,5], the resulting unions of the collections demonstrated that the identity and chemical similarity between company collections is very low. This was shown by Bayer, following its acquisition of Schering, and also by Johnson and Johnson (J&J) – finding a similar result when comparing the 3DPharmaceuticals collection with its existing compound collection. Very recently, AstraZeneca and Bayer collaborated in a precompetitive initiative to compare and analyse their entire compound collections for similarity, with a view to possible compound exchange [6]. The result was similar to the previous studies and also showed a low overlap between these two very large compound collections.

The emerging picture of a low chemical similarity between the independent screening libraries suggests that an efficient and cost-effective way to generate a highly diverse library would be to combine the libraries from multiple companies. The combined chemical space explored by all the collections would be available and this would maximise the chances of hits against a novel biological target. Merging the entire individual collections would also generate an enormous and financially 'unscreenable' library. However, combining representative subsets from the respective libraries is an alternative approach that would be expected to retain efficient exploration of chemical space and is a feasible goal.

Drug-likeness and diversity analysis

Conventional diversity analysis of multiple compound collections typically involves the exchange of chemical structures and calculation of chemical similarity scores based on, for example, 2D fingerprints [7,8], BCUT descriptors [9] or physicochemical properties. Structure exchange becomes commercially sensitive when proprietary compounds are being analysed and, when multiple companies are involved, the intellectual property issues can become complex. Comparison of compounds from proprietary collections is therefore facilitated if disclosure of the compound structure is not required. To enable this structure-free comparison, a Pipeline Pilot protocol calculating 2D fingerprints together with a large number of physicochemical parameters was developed at the University of Dundee and distributed among the pharmaceutical companies in the consortium. The protocol also assigned each compound a unique identifier. Each company ran the protocol to calculate the descriptors on an agreed number of compounds and the results were returned to Dundee for diversity analysis. This process ensured that a consistent set of descriptors was generated to allow comparison and characterisation of the proposed contributions while maintaining confidentiality. Drug-likeness can be estimated from a range of properties derived directly from the molecular structure. These include properties such as topological polar surface area (TPSA), number of rotatable bonds, molecular weight (MW) and log *P*, and also substructural features ('structural alerts') known to have assay-interfering or toxicophoric properties. A desirability index, such as the QED [10] score, can then be generated as an estimator of chemical attractiveness. All of these properties were included in the Pipeline Pilot protocol.

Diversity analysis requires further information about compound structures to be disclosed. A number of properties and descriptors encoding structural information can be directly derived from a compound structure and exchanged between owners of compound collections to enable chemical diversity calculations. Circular fingerprints, such as extended connectivity fingerprints (ECFPs), have been shown to be highly effective in encoding SAR information [7] and are well suited for estimating chemical similarity in the context of drug discovery. We chose ECFP6 fingerprints as the principal descriptor of molecular structure to maintain consistency with the pharmacological profiling of the JECL presented in this paper and these were used for Tanimoto calculations (Box 1). The seven compound collections were then combined and filtered, to reduce clusters of molecules with a Tanimoto coefficient of 1 to a single representative. This fingerprint-based deduplication could also remove very close analogues. No company collection was treated as a reference set and any could be asked to remove their compound. Once all duplicates had been removed, the final set of just over 321 000 molecules was registered in an Oracle database. A requirement for all the compounds contributed by the pharmaceutical companies was that they were not commercially available.

BOX 1

Tanimoto similarity

2D fingerprints are a representation of a molecular structure in which the structure is decomposed *in silico* into a set of substructural features. The similarity of a pair of molecules can then be calculated by counting the number of features in common compared with the features present in either molecule but not in both. This similarity score is known as Tanimoto similarity. In this analysis extended connectivity fingerprints (ECFP6) were used. These fingerprints represent the molecule as a set of substructures centred on each atom in the molecule with the substructures extending up to three bonds away from the centre atom.

Tanimoto similarity is defined as C/(A + B + C) where C = number of fingerprint features present in both molecules, A = number of fingerprint features in molecule A and not in B, and B = number of fingerprint features in molecule B and not in A.

Analysis of the JECL

The final set of selected compounds was analysed to profile drug-likeness and diversity. Diversity was assessed from a conventional chemocentric perspective using chemical diversity metrics and from a biological perspective. The spectrum of biological target activity was assessed using predicted compound activity in 1855 Bayesian activity models built using data in the ChEMBL database.

Drug-likeness and physicochemical profile

The key descriptors summarising the physicochemical properties of the library are shown in Fig. 1a–e. With the compounds all originating in the collections of pharmaceutical companies, it comes as no surprise that the collection has an attractive profile and is overwhelmingly within the Lipinski parameters for MW and log *P*. The fraction of sp3-hybridised carbon atoms (Fsp3) indicates a broad spread of hybridisation from

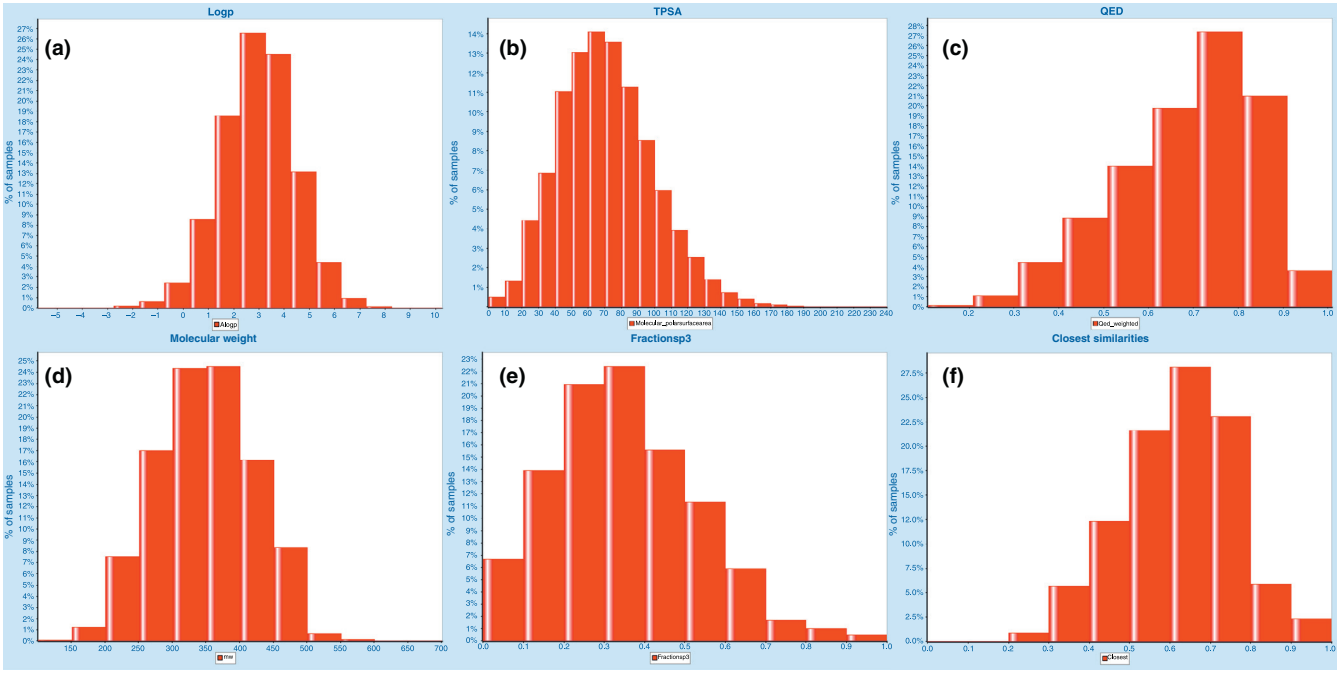

FIGURE 1

Key physicochemical parameters and diversity of the Joint European Compound Library (JECL). The key physicochemical parameters of **(a)** a log *P*, **(b)** topological polar surface area (TPSA), **(d)** molecular weight (MW), **(e)** fraction sp3 carbon atoms and **(c)** the QED desirability score are typical of a lead-like library and centred around a MW of 350 Da and a log *P* of 2–3. The fraction of sp3-hybridised carbon atoms indicates a broad spread of 3-dimensionality. **(f)** The nearest neighbour diversity of the collection using Tanimoto similarity with extended connectivity fingerprint (ECFP)6.

molecules rich in planar, aromatic moieties to very sp3-rich molecules with greater 3D characteristics.

Chemical diversity

The chemical diversity of the JECL was analysed through two different methods. First, the Tanimoto similarity of each compound with its nearest neighbour was calculated and plotted in a histogram (Fig. 1f). The profile demonstrates a diverse set with a mode Tanimoto similarity of 0.6 for each compound to its nearest neighbour. Of particular interest for the JECL is the relationship between the individual company subsets and the regions of chemical space occupied by these subsets. Do the subsets represent different areas of chemical space or are the compounds from each subset randomly distributed with respect to each other? To address this question, the same nearest neighbour diversity calculation was performed for each pairwise combination of the seven company subsets (intersubset diversity) and within each subset (intrasubset diversity). The seven intrasubset diversities define the profile of nearest neighbours within a subset whereas the 21 pairwise intersubset diversities identify the nearest neighbours of compounds in one subset in each of the other subsets. The histograms profile these nearest neighbour distributions and the results of this are shown in Fig. 2. The intrasubset diversities are shown along the leading diagonal and it is apparent that the diversity within each company subset is similar apart from in one case where the distribution of the maximum Tanimoto similarity is right- rather than left-skewed. The mode values for the intrasubset similarity are in the 0.6–0.7 range, with the exception of Co7 which has a very diverse subset with a similarity mode value at 0.3. This 0.6–0.7 similarity value is very close to that in the histogram of the entire JECL and suggests that the nearest neighbour of each JECL compound is mainly derived from the same company subset. Analysis of the intersubset histograms confirms this. The intersubset similarity histograms all have a very different profile from the intrasubset histograms. The intersubset histograms are all shifted well to the left with the most frequent nearest neighbours in the 0.2–0.3 range, demonstrating that the nearest neighbours between subsets are substantially less similar than nearest neighbours within subsets. In all cases, the intersubset diversities are all greater than the seven intrasubset diversities. Again, this is consistent with the Bayer–Schering, J&J–3DP and Bayer–AstraZeneca results. Closer comparison with the Bayer–AstraZeneca results shows that the nearest neighbour similarities of the intersubset comparisons are lower than seen in the comparison of the full Bayer–AstraZeneca collections. This is likely to be caused by the very different sizes of the compound subsets being compared and the consequent lower likelihood of a near neighbour being in the contributed subsets.

Second, the number of unique ECFP6 fingerprint features in each of the seven company subsets was calculated and normalised to the size of the subset to generate another measure of diversity. The number of unique ECFP6 features reflects the number of unique substructural features in the subset and, by normalising for the number of compounds, a measure of diversity is obtained. The fingerprint diversities of these subsets were then compared with the diversity of the JECL. To do this, the JECL was randomly subsetted into seven groups, corresponding in size to the company subsets, and the normalised unique ECFP6 feature counts were calculated for each subset. Boxplots were then generated for the seven company subsets and the seven randomly selected JECL subsets to summarise the distribution of the ECFP6 counts. These boxplots are shown in Fig. 3. The subsets generated from the combined JECL have a significantly higher number of ECFP6

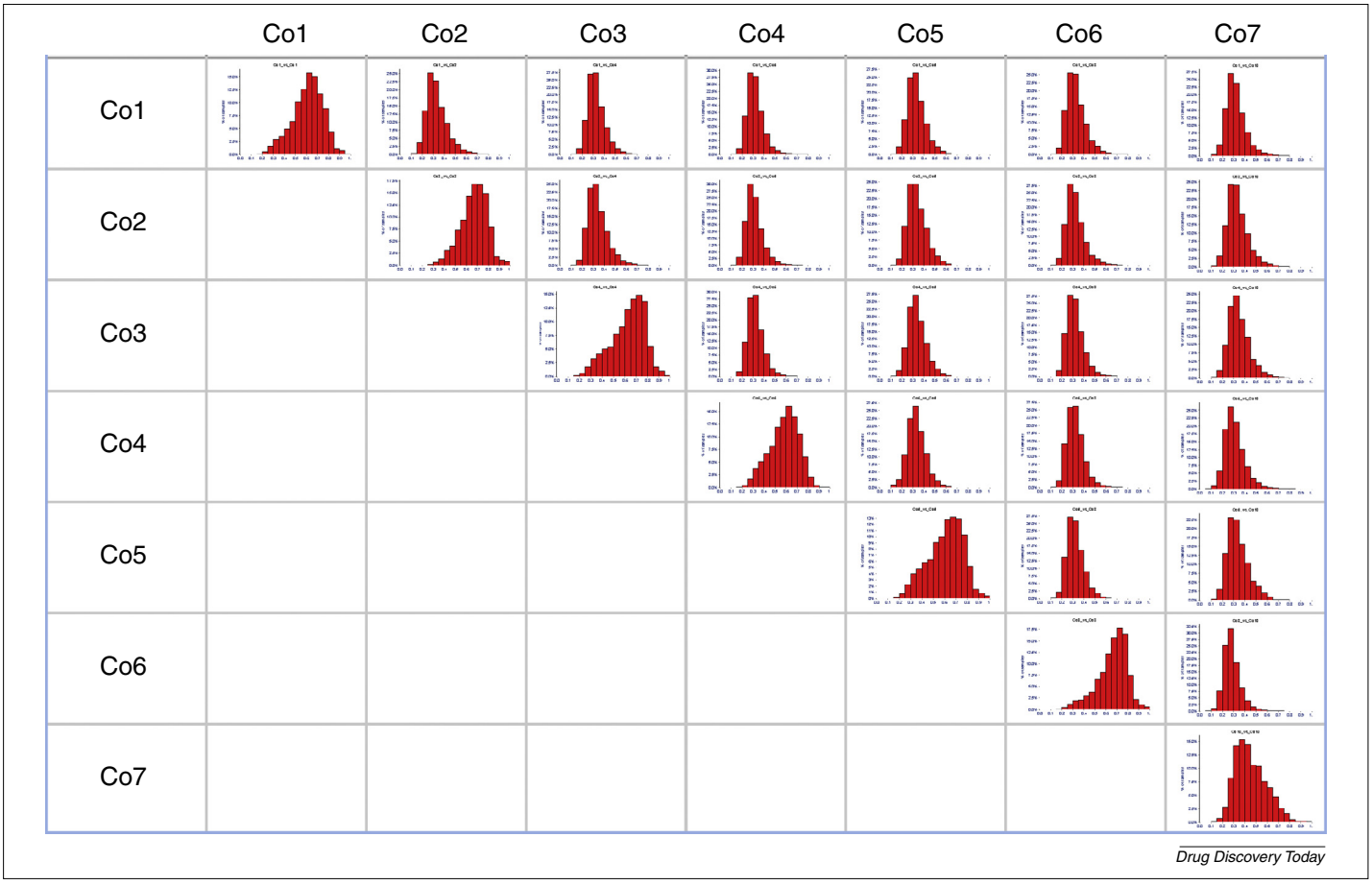

FIGURE 2

A matrix of nearest neighbour histograms showing diversity within the individual seven subsets (intrasubset diversity) along the leading diagonal; and pairwise similarities between subsets (intersubset diversity) in the off-diagonal cells. Intrasubset diversity is consistently lower than intersubset diversity.

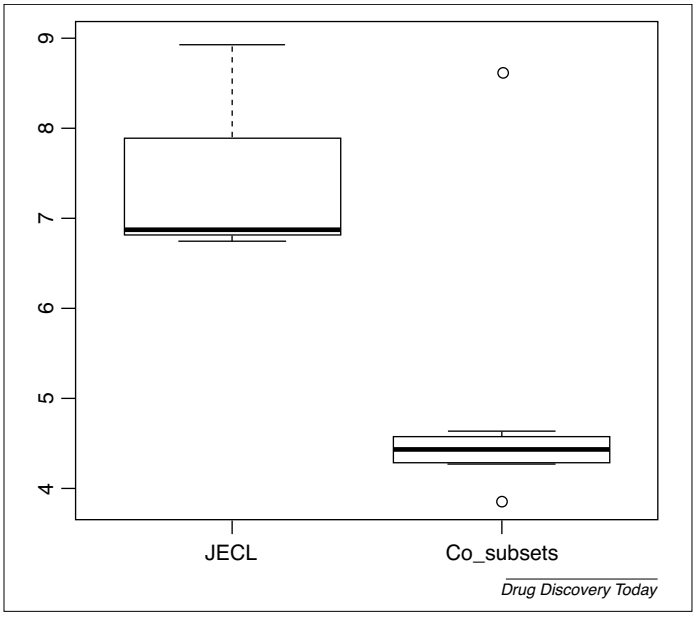

FIGURE 3

The normalised extended connectivity fingerprint ECFP6 count (y axis) in the Joint European Compound Library (JECL) and the contributing subsets. The number of normalised unique ECFP6 features in seven randomly selected subsets from the JECL is shown in the boxplot on the left. The number of normalised unique ECFP6 features in the seven individual company subsets is shown on the right. The increase in feature counts in the JECL demonstrates the effectiveness of combining subsets to increase diversity. Boxplots were generated in R.

features per compound than the individual subsets and demonstrate the efficiency of combining subsets from different sources to generate diversity.

Biological activity diversity

The key aim for the JECL is to generate hit molecules in a target-agnostic manner against a wide range of biological targets. We were therefore interested to understand the spectrum of target classes where the JECL was likely to generate biologically active molecules. Because the compounds have no publically available biological data associated with them, the predicted activity for each compound against a panel of Bayesian models was calculated. Bayesian activity models use machine learning methods and the structures of known active and inactive molecules from biological screens to derive an activity model and predict activity of compounds untested in these assays. The higher the score a compound has in a Bayesian model for a given biological target the higher is the likelihood that the compound will be active against that target. These approaches have been

TABLE 1

The level 1 (L1) classification of target classes according to ChEMBL17 and L1_L2 shows the further breakdown of this classification into subclasses.

	Number of targets	Number of bioactivities
L1_class		
Enzyme	1712	282 085
Membrane receptor	336	206 678
Ion channel	156	30 472
Transporter	59	20 426
Unclassified	424	18 800
Transcription factor	56	17 613
Cytosolic other	54	8486
Membrane other	8	1036
Structural	7	876
Secreted	27	848
Adhesion	8	570
Nuclear other	6	408
Surface antigen	12	322
L1_L2_class		
Membrane receptor_7TM1	261	187 915
Enzyme_kinase	448	98 943
Enzyme_protease	218	66 782
Enzyme_ND	833	51 932
Ion channel	156	30 472
Transporter	59	20 426
Unclassified	424	18 800
Transcription factor	56	17 613
Enzyme_reductase	40	17 294
Enzyme_hydrolase	24	12 668
Enzyme_lyase	17	11 361
Membrane receptor_ND	37	9488
Enzyme_cytochrome P450	35	8603
Cytosolic other	54	8486
Enzyme_phosphodiesterase	28	6928
Membrane receptor_7TM3	14	4540
Membrane receptor_7TM2	19	4083
Enzyme_phosphatase	35	3233
Enzyme_transferase	14	2466
Enzyme_isomerase	10	1148
Membrane other	8	1036
Structural	7	876
Secreted	27	848
Adhesion	8	570
Enzyme_NTPase	6	485
Membrane receptor_7TMFZ	1	449
Nuclear other	6	408
Surface antigen	12	322
Membrane receptor_Toll-like and interleukin (Il)-1	1	139
Enzyme_ligase	2	137
Enzyme_aminoacyltransferase	1	78
Membrane receptor_phosphatase	1	51
Enzyme_electrochemical	1	27
Membrane receptor_kinase	1	10
Membrane receptor_7TMTAS2R	1	3

demonstrated to be effective in chemogenomic analyses to predict unknown biological activity [11,12]. Bayesian activity models for 1855 diverse biological targets were built using data from the ChEMBL17 database [13,14]. Only data from published literature were used to generate the models. Compounds with fewer than four atoms and greater than 650 Da MW were excluded from model building and the threshold for activity was set to a pXC50 of 5. Targets with at least ten active compounds were retained and the predicted activity profile of each compound in the JECL against this panel of models was calculated. The top predicted targets were also mapped to their target class (e.g. kinase, ion channel or phosphatase) for global analysis of target spectrum. Leave-one-out cross-validation used in the generation of each model was used to calculate a threshold between predicted activity and inactivity by minimising false-positive and -negative rates. This threshold was used to define a minimum score for activity and, if the top score of a compound failed to meet the model threshold, it was assigned to the 'Low Score' category. This ensures that only predictions within the domain of the model are used and all compounds with low confidence or low predictions are assigned to this Low Score category.

Table 1 shows the classification by target class used in the Bayesian analysis. The classification is in two hierarchical levels derived from the ChEMBL database. The top level L1 class is very-high-level whereas the L1_L2 classification breaks this further into enzyme types. The number of biological targets and the number of experimental bioactivities associated with each target class is shown. Fig. 4 shows two pie charts demonstrating the predicted activity of the JECL in these target classes. The first pie chart shows the activity of the JECL in the very-high-level L1 target classification whereas the second shows predicted activities in the L1_L2 classification. These charts indicate a very general, target-agnostic collection suitable for screening against a wide range of target classes and with a high likelihood of generating bioactives in key pharmaceutically relevant target classes such as G-protein-coupled receptors (GPCRs), kinases and proteases.

In addition to providing a screening library, the ELF provides a full infrastructure for compound storage, screening, characterisation and hit triage. A key component of this infrastructure is the Honest Data Broker (HDB). The HDB is a secure cloud-based repository for all the compound and assay information generated by ELF projects and also provides a triaging interface to allow a final list of hit compounds to be generated. Central to this triage model is a granular set of permissions allowing carefully controlled access to the molecular structures by the triage team. These permissions maintain compound confidentiality while supporting the scientific requirements of an HTS triage. The HDB also provides functionality to reorder compounds for dose–response and further characterisation of hits such as analytical chemistry. Projects from academic and SME partners are prosecuted in close collaboration with expert biology, chemistry and computational scientists at the European Screening Centre with the final aim of building a high-quality hit list.

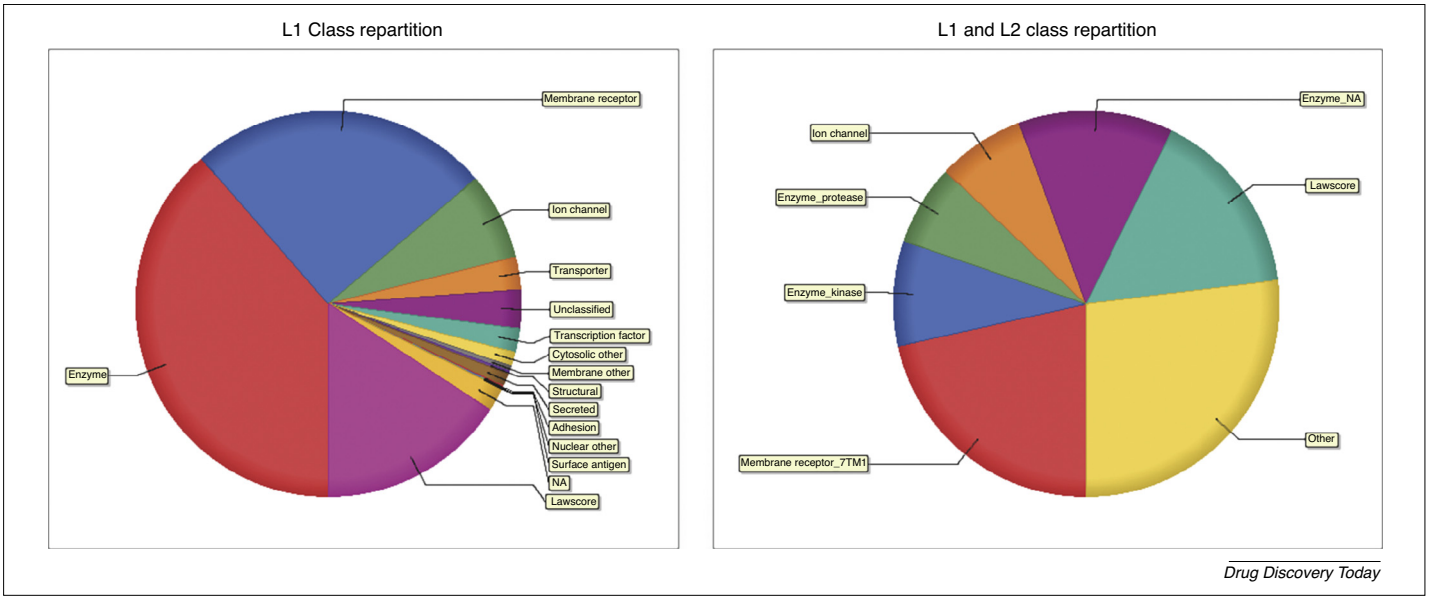

FIGURE 4

The pie charts show the predicted activities in the ChEMBL17 Bayesian models. In each chart, the top predicted targets for each compound are mapped to their target class. In the L1_L2 chart pie-chart slices of <5% of the total are merged into the 'Other' category. Compounds where highest prediction failed to meet minimum threshold in every model were assigned to the 'Low Score' category.

Concluding remarks

The JECL is a high-quality and diverse library with predicted activity against a wide range of biological targets, including classes that have been rich in therapeutically relevant targets. The value and efficiency of combining collections to generate diversity is clear from the analysis of intersubset and intrasubset diversity where the intersubset diversity demonstrates that each collection is derived from a distinct region of chemical space. It is also clear that the similarity between any pairwise combination of the seven company collections is low, and it is likely that any company seeking to increase the diversity of its screening library efficiently would benefit from a compound exchange.

The JECL is now available to academic laboratories and biotech companies for screening their targets, and expert characterisation, medicinal chemistry and modelling resource is also available to follow up the screen output. The collection will also be screened by the contributing pharmaceutical companies and provides them with a cost-effective model to explore additional chemical space beyond their own compound collections and to augment existing screening campaigns. The JECL provides a model for further precompetitive collaboration and represents a major step forwards in the level of collaboration between large pharmaceutical companies and also between commercial organisations and the European academic community. The importance of linking academic investigators with pharmaceutical screening libraries and expertise is also recognised in the recent announcement of the AstraZeneca–MRC collaboration in which up to 15 academic drug discovery programmes will be screened against the AstraZeneca compound collection, and also in the GlaxoSmithKline (GSK) Openlab Initiative. These growing partnerships will be crucial for translating the wealth of academic early-stage target validation into tangible patient benefits.

Acknowledgements

The ELF is funded with financial support from IMI JU Grant Agreement 115489. The EFPIA partners contributing compounds are AstraZeneca, Bayer, Janssen Pharmaceutica, H. Lundbeck, Merck, Sanofi-Aventis Deutschland and UCB Pharma.

References

1 Bunnage, M.E. (2011) Getting pharmaceutical R&D back on target. *Nat. Chem. Biol.* 7, 335–339
2 Mittleman, B. *et al.* (2013) Precompetitive consortia in biomedicine – how are we doing? *Nat. Biotechnol.* 31, 979–985
3 Sidders, B. *et al.* (2014) Precompetitive activity to address the biological data needs of drug discovery. *Nat. Rev. Drug Discov.* 13, 83–84
4 Schamberger, J. *et al.* (2011) Rendezvous in chemical space? Comparing the small molecule compound libraries of Bayer and Schering. *Drug Discov. Today* 16, 636–641
5 Engels, M.F. *et al.* (2006) A cluster-based strategy for assessing the overlap between large chemical libraries and its application to a recent acquisition. *J. Chem. Inf. Model.* 46, 2651–2660
6 Kogej, T. *et al.* (2013) Big pharma screening collections: more of the same or unique libraries? The AstraZeneca-Bayer Pharma AG case. *Drug Discov. Today* 18, 1014–1024
7 Rogers, D. and Hahn, M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754
8 Willett, P. (2011) Similarity searching using 2D structural fingerprints. *Methods Mol. Biol.* 672, 133–158
9 Pearlman, R. and Smith, K.M. (1998) Novel software tools for chemical diversity. *Persp. Drug Discov. Design* 9–11, 339–353
10 Bickerton, G.R. *et al.* (2012) Quantifying the chemical beauty of drugs. *Nat. Chem.* 4, 90–98
11 Glick, M. *et al.* (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* 46, 1124–1133
12 Martinez-Jimenez, F. *et al.* (2013) Target prediction for an open access set of compounds active against *Mycobacterium tuberculosis*. *PLoS Comput. Biol.* 9, e1003253
13 Gaulton, A. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107
14 Bento, A.P. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42, D1083–D1090

Jérémy Besnard[1]
Philip S. Jones[2]
Andrew L. Hopkins[1]
Andrew D. Pannifer[2]

[1]College of Life Sciences, University of Dundee, Dow Street, Dundee, DD1 5EH Scotland, UK
[2]College of Life Sciences, University of Dundee, Biocity Scotland, Bo'ness Road, Newhouse, ML1 5UH Scotland, UK